

- 1 -

Express Mail Label No. EV215731041US  
Date: 8-1-03

## PERSISTENT RESERVATION IO BARRIERS

5

## RELATED APPLICATION

This application is a continuation of U.S. Application No. 09/546,806, filed April 11, 2000. The entire teachings of the above application are incorporated herein by reference.

10

## BACKGROUND OF THE INVENTION

A cluster is a plurality of nodes physically connected to an inter-node communication network. Each of the cluster nodes is a computer system. The computer system may include a Central Processing Unit ("CPU"), memory, an inter-node communications interface and IO subsystem.

15

A storage device may be connected to the IO subsystem in a node. The storage device may be shared by a plurality of nodes by connecting the device to the IO subsystem in each node. By sharing the storage device amongst a plurality of nodes, multiple paths are provided for accessing the storage device. The multiple paths to the storage device provide redundancy in the case of a failure in one of the nodes by sending an IO request to the storage device through a non-failed node.

20

A well-known standard interface for connecting storage devices to an IO subsystem is the American National Standards Institute ("ANSI"). Small Computer System Interface ("SCSI"). ANSI SCSI defines a protocol for accessing storage devices connected to a storage network. The SCSI protocol permits a storage device connected to a storage network to be shared by a plurality of nodes. The IO subsystem includes in each node a storage network controller. The storage network controller includes logic for issuing IO commands over the storage network storage device. The IO commands include a command to read data from the storage device and a command to write data to the storage device.

25  
30

ANSI SCSI includes a Persistent Reserve command. The Persistent Reserve command allows a storage device to be shared by more than one cluster node. Each

storage network controller issues a Persistent Reserve command to the storage device to register with the storage device. A second Persistent Reserve command is issued to reserve the device by specifying the access type. The storage device stores a list of registered storage network controllers with a corresponding registration key  
5 and the type of access permitted.

The Persistent Reserve command provides security by requiring registered storage network controllers to provide their registration key before allowing the storage network controller to perform commands restricted to members of the group of registered storage network controllers. For example, if each storage network  
10 controller registers with registration type "write exclusive registrants only", only registered storage network controllers have permission to write to the storage device but all other storage network controllers have permission to read from the storage device.

In a cluster, a node failure is communicated to survivor nodes on the inter-  
15 node communication network. Upon detecting the node failure, access to the storage device may be provided on an alternative path through survivor node in the cluster connected to the storage device. However, before access can be provided on the alternative path, all the pending IO commands issued by the failed node must be completed or aborted in the storage device in order to guarantee that these IO  
20 commands do not interfere with future IO commands from surviving cluster members. A survivor node in the cluster issues a Persistent Reserve command to the shared storage device to request the completion or abortion of all IO commands issued by the failed node in the cluster.

There are two types of SCSI physical connections. A parallel SCSI physical  
25 connection provides for the connection of a maximum of sixteen devices including storage devices and storage network controllers. A serial SCSI physical connection provides for the connection of 264 devices including storage devices and storage network controllers, switches and routers. Through the SCSI physical connection, a cluster storage device may be accessed by several cluster nodes; that is, nodes  
30 connected to a cluster and non-cluster nodes. Through the use of the Persistent Reservation command write access to a cluster storage device can be limited to

registered cluster nodes by registering each cluster node with “write exclusive registrants only” registration type.

The “write exclusive registrants only” state remains in effect as long as one of the cluster nodes is registered with the storage device. However, if the persistent  
5 reservation from the last cluster node is removed, a non-cluster node or a cluster node from another cluster may write to the storage device and corrupt data stored in the storage device.

#### SUMMARY OF THE INVENTION

10 The present invention provides a method for sharing a storage device amongst a plurality of computers while providing data integrity in the storage device. A computer is registered with the storage device by storing a computer identifier associated with a reserved access type in the storage device. Access to the storage device is provided to the registered computer dependent on the registered  
15 computer’s stored identifier. The type of access provided to the registered computer is dependent on the stored access type. Upon loss of knowledge of the stored identifier in the shared storage device by the registered computer, the identifier for the computer stored in the shared storage device is replaced with a new identifier for the registered computer. The registered computer may be a currently registered  
20 computer or a previously registered computer.

Upon detecting a failure in one of the registered computers, one of the survivor registered computers removes the registration for the detected failed computer by requesting deletion of the identifier associated with the reserved access type for the detected failed registered computer in the shared storage device.  
25 Outstanding commands in progress from the detected failed registered computer to the shared storage device are aborted. All commands to the shared storage are stalled until all pending commands issued by the detected failed computer are aborted.

The identifier for each computer may be unique or the identifier may be  
30 initialized to the same value. If the identifier is initialized to the same value, the

identifier assigned to another computer stored after detection of the failed node differs from the previously stored identifiers.

#### BRIEF DESCRIPTION OF THE DRAWINGS

5

The foregoing and other objects, features and advantages of the invention will be apparent from the following more particular description of preferred embodiments of the invention, as illustrated in the accompanying drawings in which like reference characters refer to the same parts throughout the different views. The drawings are not necessarily to scale, emphasis instead being placed upon illustrating the principles of the invention.

10

Fig. 1 is a block diagram illustrating a cluster with two nodes, a non-cluster node and a storage device connected to a storage network according to the principles of the present invention;

15

Fig. 2 is a block diagram illustrating a preferred embodiment for the cluster shown in Fig. 1;

Fig. 3 is a block diagram illustrating storage network routines and shared device routines stored in the memory system of each cluster node shown in Fig. 2 for managing the shared storage device;

20

Fig. 4A is a block diagram illustrating a SCSI Persistent Reserve Out command;

Fig. 4B is a block diagram illustrating the parameters communicated in the SCSI Persistent Reserve Out command shown in Fig. 4A;

Fig. 4C is a block diagram illustrating a SCSI Persistent Reserve In command;

25

Fig. 5 is a block diagram illustrating a Persistent Reserve table stored in the shared storage device;

Fig. 6 is a flow chart illustrating the steps executed in the memory system shown in Fig. 2 for managing access to the shared storage device.

30

#### DETAILED DESCRIPTION OF THE INVENTION

A description of preferred embodiments of the invention follows.

Fig. 1 illustrates a cluster 100 with two cluster nodes 102a and 102b, a non-cluster node 110 and a storage device 104 connected to a storage network 108. The

cluster nodes 102a and 102b are physically connected through an inter-node communications bus 106. Each of the cluster nodes 102a and 102b and the non-cluster node 110 are physically connected to the storage device 104 through the storage network 108.

5           The storage device 104 may be a disk storage device, a tape storage device, a Redundant Array of Independent Disks ("RAID") or any other type of device well known in the art which can be physically connected to a storage network 108. Each cluster node 102a and 102b and the non-cluster node 110 may access the storage device 104.

10           The storage network 108 may be a serial storage network such as, Fibre Channel and may include Fibre Channel switches and routers. Alternatively, the storage network may be a parallel storage network. The serial or parallel storage network 108 may be the American National Standards Institute ("ANSI") Small Computer Systems Interface ("SCSI") storage network. Each of the cluster nodes  
15 102a, 102b, the non-cluster node 110 and the storage device 104 include one or more storage network controllers 112 for communicating over the storage network 108.

Each storage network controller 112 includes a unique identifier identifying the storage network controller 112 on the storage network 108. The storage device  
20 104 may be reserved for exclusive use by one or more storage network controllers 112 by storing the storage network controller's unique identifier with an associated access privilege for the storage network controller 112 in the storage device 104. For example, storage network controllers 112 located in a cluster node 102a or 102b may have read and write access privilege for the storage device 104 and a storage  
25 network controller 112 in the computer system 110 may have read-only access privilege for the storage device 104. By providing access privileges, a storage device 104 can be reserved for exclusive write access by cluster nodes 102a and 102b in a cluster 100. Thus, even though the non-cluster system node 110 is physically connected to the storage device 104 through the storage network 108, a  
30 storage network controller 112 connected to the storage network 108 may not have permission to write to the storage device 104.

Fig. 2 is a block diagram illustrating a preferred embodiment for a cluster 100 with two cluster nodes 102a and 102b. The cluster 100 is not limited to the two cluster nodes 102a and 102b shown. It is well known in the art that a cluster 100 may include more than two nodes 102a and 102b. Four and eight node clusters are well known in the art. For example, Compaq Computer Corporation's TruCluster64 UNIX 4.0 and TruCluster64 UNIX 5.0 provide support for four and eight node clusters.

Each cluster node 102a, 102b includes a respective CPU module 200a, 200b; a system bus interface 204a, 204b, a storage network controller 112a, 112b, an inter-  
10 node communication memory 202a, 202b and a memory system 208a, 208b.

The inter-node communication memory 202a, 202b provides communication between cluster nodes 102a and 102b in the cluster 100 over the inter-node communication bus 106. The inter-node communications bus 106 is used to communicate configuration parameters such as shared storage devices and to detect  
15 a node failure. For example, a survivor node may detect a failed node if it does not receive a response to a keep-alive command forwarded over the inter-node communications bus 106 to the failed cluster node 102a, 102b.

Thus, multiple paths are provided to the storage device 104 for access by a client computer system (not shown) connected to the cluster 100. In the case of a  
20 failure of a component in one of the cluster nodes 102a, 102b, access to the storage device 104 is provided through the survivor cluster node 102a, 102b. The CPU module 200 in each node 102a and 102b executes storage network routines (not shown) and shared device routines (not shown) stored in each memory system 208.

Each storage network controller 112a, 112b communicates with storage  
25 devices including storage device 104 connected to the storage network 108 by forwarding a storage network command on the storage network 108. The storage network command includes a unique storage network identifier for the storage network controller 112 and a unique storage network identifier for the storage device 104.

30 Fig. 3 is a block diagram illustrating storage network routines and storage device routines stored in the memory system 208 of each cluster node 102a, 102b shown in Fig. 2 for managing the storage device 104. The software modules stored

in each memory system 208a, 208b include a connection manager 302 and a device Input/Output ('I/O') subsystem 311.

The connection manager 302 includes a node identifier manager 310. The node identifier manager 310 assigns a key (an identifier) to a cluster node 102a, 102b each time the cluster node 102a, 102b joins the cluster 100. The key used to register a cluster node 102a, 102b with a storage device 104 may be the same number for each cluster node 102a, 102b in the cluster 100 or may be the unique node identifier generated by the connection manager for each cluster node 102a, 102b in the cluster 100.

10        If the key used is the same number for each cluster node 102a, 102b, the key is incremented for each cluster node 102a, 102b remaining in the cluster 100 after the detection of a failed cluster node 102a, 102b or a previously failed cluster node 102a, 102b rejoins the cluster 100.

15        If the key assigned to a cluster node 102a, 102b is a unique number for each cluster node 102a, 102b, the survivor cluster node 102a, 102b retains the previously assigned key after the detection of a failed cluster node 102a, 102b or a previously failed cluster node 102a, 102b rejoins the cluster. However, upon rejoining the cluster, a previously failed cluster node 102a, 102b receives a new key. Thus, a cluster node 102a, 102b is not required to remember the key that was previously  
20 assigned by the connection manager 302.

For example, the identifier can include a cluster node number and an incarnation number. If the cluster node number assigned to cluster node 102a is 1 and the cluster node number assigned to cluster node 102b is 2, on the first registration the identifier for cluster node 102a is set to 11. If the incarnation number is incremented each time a cluster node 102a, 102b re-registers with the storage device 104, on the second registration the identifier for cluster node 102a is set to 21  
25

The device IO subsystem 311 includes a device IO request dispatcher 304, an IO driver 306 and a host bus adapter driver 308. The device IO request dispatcher 304 includes a shared device manager 312. The shared device manager 312 includes  
30 a register routine 318, and an unregister routine 322. The unregister routine 322 includes a stall routine 324 and a pre-empt and abort routine 326. The shared device manager 312 sends storage network commands to the storage device 104 in the

cluster 100 and receives responses from the storage device 104. The storage network commands include a register key command and a reserve command generated in the register routine 318, and a read keys command and a preempt and abort command generated in the unregister routine 322.

5           The register key command registers the specified key for the cluster node 102a, 102b with the storage device 104. The reserve command requests that the access mode for the device be set to “write exclusive read only”. The read keys command requests that the shared device manager 312 return a list of the cluster nodes 102a, 102b that have previously registered with the storage device 104 by  
10       returning a list of registration keys stored in the storage device 104. The preempt and abort command remaps registrations, reservations, and aborts IO commands.

          The IO driver 306 in the device IO subsystem 311 prepares register key commands and the read key commands to be forwarded on the storage network 108. In a storage network 108 with a Fibre Channel or parallel SCSI physical connection,  
15       the storage network command format is defined in the SCSI command protocol. The SCSI commands are described in conjunction with Figs. 4A-4C.

          If the storage network 108 is a SCSI network, the prepare SCSI commands routine 314 in the IO driver 306 prepares SCSI commands for the register key command, reserve command, preempt and abort command and the read keys  
20       command. After the IO driver 306 has prepared a SCSI command to be forwarded on the storage network 108 to the shared storage device 104, the issue SCSI command routine 316 in the host bus adapter driver 308 forwards the SCSI command through a storage network controller 112. The storage network controller 112 physically forwards the SCSI command on the storage network 108 to the  
25       storage device 104. Each host bus adapter driver 308 may communicate with a plurality of storage network controllers 112. Typically a host bus adapter driver 308 communicates with a particular type of storage network controller 112.

          Fig. 4A is a block diagram illustrating a SCSI Persistent Reserve Out command 400. Upon a request from the register routine 318, the host bus adapter  
30       driver 308 forwards a SCSI Persistent Reserve Out command 400 to the storage device 104, to register a cluster node 102a, 102b with the storage device 104 and another Persistent Reserve Out command 400 to reserve the device for “write

exclusive read only” access. Upon a request from the unregister routine 322, after detection of a failed registered cluster node 102a, 102b by the connection manager 302, the host bus adapter driver 308 forwards a SCSI Persistent Reserve Out command 400 to the storage device 104, to preempt and abort outstanding storage network commands issued from the failed registered cluster node 102a, 102b.

The SCSI Persistent Reserve Out command 400 includes a Persistent Reserve operation code 402, a service action field 406, a scope field 408, a type field 410, parameter list length fields 420, 422, a control field 424 and reserved fields 404, 412, 414, 416, 418. The contents of the service action field 406 determine the action to be performed by the storage device 104.

The contents of the scope field 408 determine whether the key is to be registered with the entire shared storage device 104 or a media changer element if the shared storage device is a media changer.

The content of the type field 410 determines the type of access assigned to the cluster node 102a, 102b associated with the registration key. In the cluster 100, each cluster node 102a, 102b registers with the type field 410 set to “write exclusive registrants only.” While a storage device 104 is reserved with type field 410 set to “write exclusive registrants only,” only registered cluster nodes 102a, 102b can issue “write” type commands to the shared storage device 104. The cluster node’s assigned key is forwarded in a SCSI Persistent Reserve command’s parameter list. The parameter list is described in conjunction with Fig. 4B.

Fig. 4B is a block diagram illustrating the parameter list 436 forwarded with the Persistent Reserve Out command 400 shown in Fig. 4A. The parameter list length fields 420, 422 (Fig. 4A) in the Persistent Reserve Out command 400 store the total number of bytes in the parameter list 436. The parameter list 436 includes a reservation key 438, a service action reservation key 440, a scope specific address 442, reserved fields 444, 446 and an obsolete field 448.

To register a cluster node 102a, 102b, the prepare SCSI command routine 314 (Fig. 3) prepares two Persistent Reserve Out commands 400 to be issued to the storage device 104. In the first Persistent Reserve Out command 400, the prepare SCSI command routine 314 stores the key assigned to the cluster node 102a, 102b by the DeviceIO Request Dispatcher 304 (Fig. 3) in the service action reservation

key field 440 and sets the service action field 406 to “register and ignore existing key” and forwards the Persistent Reserve Out command 400 and the parameter list 436 to the host bus adapter driver 308 (Fig. 3). In the second Persistent Reserve Out command 400, the prepare SCSI command routine 314 sets the service action field 406 to “reserve” and the type field 410 to “write exclusive registrants only” and forwards the Persistent Reserve Out command 400 and the parameter list 436 to the host bus adapter driver 308 (Fig. 3).

Upon detecting a failed cluster node 102a, 102b, a surviving cluster node 102a, 102b issues a Persistent Reserve Out command 400 with service action field 406 set to “preempt and abort,” the key assigned to the survivor cluster node 102a, 102b is stored in the reservation key field 438 and the key assigned to the failed cluster node 102a, 102b is stored in the service action reservation key field 440. The scope specific address field 442 is set to ‘0’ unless the reserve request is directed to an element in a media changer.

Fig. 4C is a block diagram illustrating a Persistent Reserve In command 426. The Persistent Reserve In command 426 includes a Persistent Reserve In operation code field 428, a service action field 430, allocation length fields 432, 434, a control field 424 and reserved fields 412, 414, 416, 418, 420. Setting the service action field 430 to “read keys” or “read reservations” returns a list of registered keys and associated data stored in the storage device 104. The associated data includes the type field 410 and scope field 408 written to the storage device in the Persistent Reservation Out command 400.

Fig. 5 is a block diagram illustrating a Persistent Reserve table 500 stored in the storage device 104. The Persistent Reserve table 500 includes a reservation entry 502 for each reservation. The reservation entry 502 includes an initiator identifier 512 and a reservation descriptor 514. The initiator identifier 512 identifies the storage network controller 112 from which the Persistent Reserve Out command 400 was received. Each storage network controller 112 on a storage network 108 has a unique identifier.

The reservation descriptor 514 includes a reservation key field 504, a reservation type field 506, a reservation scope field 508 and a scope specific address field 510. The data stored in the reservation descriptor 514 is dependent on the

contents of the parameter list 436 forwarded with the Persistent Reserve Out command 400. The contents of the reservation descriptor 514 are returned in the reservation list (not shown) with the Persistent Reserve In command 426.

The reservation key field 504 stores a reservation key assigned to a cluster node 102a, 102b by the DeviceIO Request Dispatcher 304 (Fig. 3). Each cluster node 102a, 102b may include a plurality of storage network controllers 112. The storage device 104 is protected from non-registered storage network controllers 112 while a storage network controller 112 is registered with reservation type set to “write exclusive registrants only”. The storage device 104 processes a write command received from a cluster node 102a, 102b if the persistent reservation table 500 indicates that there are no current persistent reservations. The shared storage device 104 searches the persistent reservation table 500. If the storage device 104 finds a reservation entry 502 in the persistent reservation table 500 for the initiator identifier 512 from which the write command was received, the write command is processed. However, if there is no reservation entry 502 for the initiator identifier 512 and there is a reservation descriptor 514 with the reservation type field 506 set to “write exclusive registrants only,” a non-registered storage network controller 112 cannot write to the storage device 104.

Fig. 6 is a flow chart illustrating the steps executed in the memory system 208 shown in Fig. 2 for managing access to the shared storage device 104. Fig. 6 is described in conjunction with Figs. 2, 3, 4A-C and 5.

At step 600, the DeviceIO Request Dispatcher 304 determines whether to register a cluster node 102a, 102b in the cluster 100 with the shared storage device 104 dependent on whether the DeviceIO Request Dispatcher 304 can see the device from the storage network controller 112a, 112b. If so, processing continues with step 602. If not, processing continues with step 608.

At step 602, the cluster node 102a, 102b determines whether it is the first cluster node 102a, 102b in the cluster 100 to register with the storage device 104. If so, processing continues with step 604. If not, processing continues with step 606.

At step 604, the DeviceIO Request Dispatcher 304 assigns a key, for the cluster node 102a, 102b. The key may be either cluster global or node specific from the connection manager 302. The register routine 318 in the device IO request

dispatcher 304 issues a request to register the cluster node 102a, 102b with each desired shared storage device 104 connected to the storage network 108. The register request is translated to a SCSI Persistent Reserve Out command 400 in the prepare SCSI commands routine 314 in the IO driver 306 and issued to the shared storage device 104 in the issue SCSI commands routine 316 in the host bus adapter driver 308. In particular, first the SCSI Persistent Reserve Out command 400 with “register and ignore” in the service action field 406 and assigned key in the parameter list (not shown) is issued, then the SCSI Persistent Reserve Out command 400 with “write exclusive registrants only” in the type field 410 and “reserve” in the service action field 406 is issued. The first cluster node also removes registrations for all other registered cluster nodes in the storage device 104 as will be discussed later in conjunction with step 616. Processing continues with step 608.

At step 606, a cluster node 102a, 102b in the cluster node 100 has already registered. Thus, a reservation already exists for the cluster 100 in the storage device 104. The DeviceIO Request Dispatcher 304 assigns a key, for the cluster node 102a, 102b. The key may be either cluster global or node specific from the connection manager 302. The register routine 318 in the device IO request dispatcher 304 issues a request to register the cluster node 102a, 102b with each desired shared storage device 104 connected to the storage network 108. The register request is translated to a SCSI Persistent Reserve Out command in the prepare SCSI commands routine 314 in the IO driver 306 and issued to the shared storage device 104 in the issue SCSI commands routine 316 in the host bus adapter driver 308. Only the SCSI Persistent Reserve Out command 400 with “register and ignore” in the service action field 406 and assigned key in the parameter list (not shown) is issued. Processing continues with step 608.

At step 608, the connection manager 302 through the inter-node communications memory 202 monitors communication between cluster nodes 102a, 102b in the cluster 100. Upon detecting a request from a node to join the cluster 100, processing continues with step 608. The request to join the cluster 100 may be from an existing previously failed cluster node 102a, 102b or from a node requesting membership of the cluster for the first time. If a request to join the cluster 100 is not detected processing continues with step 602.

At step 608, access is enabled to the shared storage device 104. The registered cluster nodes 102a, 102b having gone through step 600 are permitted to write data to the shared storage device because the type of reservation enabled is “write exclusive registrants only”. This type was stored in the reservation type field 506 in the reservation entry 502 in the persistent reservation table 500 in response to the issued SCSI Persistent Reserve Out command 400 with type field 410 set to “write exclusive registrants only” in step 604. Processing continues with step 610.

At step 610, the connection manager 302 through the inter-node communication memory 202 monitors communication between cluster nodes 102a, 102b in the cluster 100. Upon detecting a failed node 102a, 102b, processing continues with step 616. If a failed cluster node 102a, 102b is not detected, processing continues with step 612.

At step 612, the connection manager 302 determines if there is a request to perform an orderly shutdown of the operating system. If so, processing continues with step 614. If not, processing continues with step 602.

At step 614, a shutdown routine (not shown) in the cluster node 102a, 102b performs an orderly shutdown of the operating system so that no data corruption occurs. The system may also be shutdown after the detection of an illegal condition; that is, a “system crash”. Typically a system crash recovery routine is executed after the detection of an illegal condition. The system crash recovery routine attempts to write the state of the operating system upon detection of the illegal condition to a storage device before executing the shutdown routine.

The shutdown routine may return IO requests queued in the shared storage device 104 but does not remove the reservation or registration for the cluster node 102a, 102b from the shared storage device 104. As long as one reservation remains on the shared storage device 104 for one cluster node 102a, 102b, only registered cluster nodes are permitted to write to the shared device. Upon restarting the cluster node 102a, 102b processing begins at step 600.

At step 616, the unregister routine 322 in the survivor cluster node 102a, 102b requests the registered keys from the shared storage devices 104 associated with the failed node. Upon a request from the unregister routine 322 executing in the survivor node 102a, 102b to read the keys, the prepare SCSI commands routine

314 prepares a Persistent Reserve In command 426 (Fig. 4C) with the service action field 430 set to "read keys". The issue SCSI commands routine 316 in the host bus adapter driver 308 issues the Persistent Reserve In command to the shared storage device 104. The shared storage device 104 returns a list of registration keys.

5           The unregister routine 322 in the survivor node 102a, 102b then removes the failed cluster node's registration. The abort routine 326 in the survivor node 102a, 102b requests that all queued storage network commands from the failed cluster node 102a, 102b be returned. Upon a request to remove the registration and reservation for a failed cluster node 102a, 102b from unregister routine 322  
10       executing in the survivor cluster node 102a, 102b, the prepare SCSI commands routine 314 prepares a Persistent Reservation Out command 400 (Fig. 4A) with the service action field 406 set to "preempt and abort" and the reservation key field 438 in the parameter list 436 (Fig. 4B) set to the survivor node's key and the service action reservation key field 440 set to the failed node's key. The issue SCSI  
15       command routine 316 in the host bus adapter driver 308 forwards the Persistent Reservation Out command 400 to the shared storage device 104.

          Upon detection of the failed node 102a, 102b, the stall routine 324 in the survivor node 102a, 102b stalls all queued storage network commands issued to the shared storage device 104. Upon receiving a request to abort from the abort routine  
20       326 in the survivor node 102a, 102b, the queued storage network commands issued by the failed cluster node 102a, 102b are aborted and the reservation entry and the registration 502 for the failed cluster node 102a, 102b are removed from the persistent reservation table 500. Processing continues with step 600.

          Thus, the shared storage device 104 is protected from data corruption by  
25       other devices connected to the storage network 108 even when all the cluster nodes 102a, 102b are shutdown. Also, as described in conjunction with step 606 a cluster node 102a, 102b does not need to remember the key with which it registered in order to rejoin the cluster 100 after all cluster nodes 102a, 102b in the cluster 100 have been shutdown.

30           It will be apparent to those of ordinary skill in the art that methods involved in the present invention may be embodied in a computer program product that includes a computer usable medium. For example, such a computer usable medium

may consist of a read only memory device, such as a CD ROM disk or conventional ROM devices, or a random access memory, such as a hard drive device or a computer diskette, having a computer readable program code stored thereon.

While this invention has been particularly shown and described with  
5 references to preferred embodiments thereof, it will be understood by those skilled in the art that various changes in form and details may be made therein without departing from the scope of the invention as defined by the appended claims.